DOCUMENT RESUME

ED 200 216 IR 009 250

AUTHOR Buckland, Lawrence F.
TITLE Data Input for Libraries: State-of-the-Art Report.
INSTITUTION Inforonics, Littleton, Mass.
PUB DATE 1 Aug 80
NOTE 9p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cataloging; *Cost Effectiveness; *Input Output Devices; Library Catalogs; Library Technical Processes; *Minicomputers; *Online Systems; Optical Scanners
IDENTIFIERS Word Processing

ABSTRACT

This brief overview of new manuscript preparation methods which allow authors and editors to set their own type discusses the advantages and disadvantages of optical character recognition (OCR), microcomputers and personal computers, minicomputers, and word processors for editing and database entry. Potential library applications are also indicated, including such special problems as converting back files, and available commercial services are mentioned. (RAA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED200216

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

STATE-OF-THE-ART REPORT

DATA INPUT FOR LIBRARIES

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Lawrence F. Buckland

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Lawrence F. Buckland
President
Inforonics, Inc.

1 August 1980

IR009250

STATE-OF-THE-ART REPORT

DATA INPUT FOR LIBRARIES: Techniques for catalogers, bibliographers, information specialists and reference librarians.

INTRODUCTION

Publishers are beginning to adopt new methods of preparing manuscript which allows authors and editors to "set their own type". These new techniques are directly applicable to librarians for creating their catalog data bases. These methods use word processors, low cost micro-computers, intelligent data entry terminals, on-line computers, or Optical Character Recognition (OCR) equipment to capture cataloging on machine media. The obvious advantage of the new method is reduced cost because keyboarding does not have to be repeated by a data processing keyboard operator.

In its 18 years of existence, Inforonics has seen a shift in the customer-vendor interface toward in-house preparation of data files. A significant amount of the input, keyboarding and data editing is performed by our customers at their location. This transition has not been without its problems, but we have learned a number of lessons and have developed a set of successful keyboard entry and edit procedures which customers use to enter data. The choice of the best input method depends on the size and number of jobs a customer processes. The purpose of this report is to provide our present and potential customers with an overview of what is possible by briefly describing each procedure and providing an estimate of processing costs. This report is intended to serve as a guide in selecting the most suitable input method. A final section of the report deals with special procedures for encoding backlogs of data when converting a large data file to a mechanized system.

PROCEDURES AND EQUIPMENT FOR DATA ENTRY OF TEXT

The following sections describe data entry procedures and data editing procedures available from Inforonics. They are arranged in order of increasing equipment cost.

1. Optical Character Recognition (OCR)

Although Inforonics does not accept OCR copy from customers, it does maintain an OCR installation for its own use. The procedure is described here to make the scope of the report complete.

The data input system which requires the simplest and least expensive equipment is optical character recognition. Using this technique, data is typed in a standard format on sheets of smooth surface paper. These sheets are fed into the input hopper of an OCR scanner which

converts them to computer form. Data is output on computer storage media, directly into a computer, or to a computer via a telecommunication link. A sample OCR form for a catalog record can be obtained by writing Inforonics, Inc.

The cost of the typewriter suitable for keyboarding OCR sheets is slightly less than $1,000.00 and the average industry cost for scanning is approximately $0.30 per 1,000 characters for clean OCR sheets. (Approximately $.18 for data for a catalog record.)

Some scanners have editing capability which allows an editor to correct OCR sheets by marking. The scanner interprets the marks, making the specified corrections. In this process, an operator monitors the scanning process by correcting data which the scanner can't read and inserting data where necessary. Costs for this monitored type of scanning are considerably higher and can range up to $0.80 per 1,000 characters.

Although OCR scanning offers the advantage of the lowest equipment cost, significant difficulties can occur in production of computer media this way. These stem from possible scanner inaccuracies in recognizing the keyboarded characters on the OCR sheets. There are numerous reasons for errors, including improper clarity of type impression, ribbon and paper used, cleanliness of the scanner, alignment of the typescript on the OCR sheet, etc. It suffices to say that to be successful, all components of the OCR process have to be "finely tuned". This can be done easily if all components are in one location. Inforonics operates such a system for customers for whom we do data entry. However, our experience has shown that the OCR sheet quality control of work done off the premises cannot be maintained adequately and therefore we do not accept this type of customer input.

2. Micro-computers and Home Computers

Prices are rapidly decreasing for computers assembled from large scale micro-circuitry. These computers are now priced from $500.00 to $3,000.00. One sizeable share of this market is the so-called "home computer" which can be used as a data input device. Low cost programs are available for home computers which can be used to create text files. Inforonics presently accepts diskettes from the Radio Shack TRS-80 model computer which is available for approximately $1,500.00 without a printer and approximately $2,250.00 with a printer. Diskettes are delivered from a customer's TRS-80 to Inforonics where they are read and transmitted to our computer. This service is new but promises to be an important method of customer data entry. The specification for an ASCII file recorded on the TRS-80 diskette can be obtained by writing Inforonics, Inc. The cost for data entry of this type is approximately $0.15 per 1,000 characters ($0.08 for input data to create a catalog record). Its equipment cost is less than word processors and mini-computers and it is a more reliable input method than OCR.

The one caution that must be mentioned is that the software support for home computers may not be adequate for the librarian. Although

there is software available for these machines, it is a mass market product and is not necessarily directly applicable to a library's input problem. Inforonics and other text processing service companies accept disks to a local standard specification, however, they leave the software setup needed to produce these disks up to the customer and the home computer "store". If data recording capability problems arise, Inforonics will solve the problem by providing software services on a consulting basis. Because such software consulting services usually require costly programming, its cost can approach or exceed that of the home computer if one is not careful.

3. Word Processors

The increasing availability of word processing in the office environment offers the librarian another option for keyboarding text for catalog records. Word processors are micro- or mini-computers with special purpose software for keyboarding, editing and printing text. These machines are used to produce all types of office documents ranging from correspondence to manuals and reports. More recently word processor manufacturers are expanding their functions to simple accounting and file search so that they are becoming "office automation" machines. Costs range from approximately $10,000.00 for a single terminal system to $30,000.00 - $40,000.00 for a multi-terminal shared logic system.

Inforonics accepts data from word processors and enters it into its TPS processing system at a cost of $0.15 per 1,000 characters. The primary advantages of word processing are equipment availability and superior editing capability. However, problems with word processing occur because of incompatibility of the storage diskettes used in different brands of word processors. This diskette incompatibility could mean that a typesetting service vendor such as Inforonics must have several types of diskette drives and many different sets of disk reading software to accept and read all types of diskettes. At Inforonics, this approach was thought too awkward and too expensive. We have chosen to standardize input by accepting word processing data via telecommunication. This requires that the word processor have transmission capability with one of the industry standard communication protocols and that the publisher use a standard printwheel character set.

Currently, Inforonics accepts TTY and BISYNC protocols. The TTY protocol which is most common on word processing machines, does not have error checking beyond parity. Thus, the telephone transmission must be reliable or errors will creep into the data. The long range solution to reliable word processing transmission is to use error correcting methods, and users, when selecting a word processor, should question the manufacturer on this point. (For an account of one customer's use of a word processor for text input see Lawrence F. Buckland, "Book Composition by Authors--System and Equipment Considerations." In the 1979 SSP Proceedings, pp. 49-53 Washington: Society for Scholarly Publishing, 1980. Reprint available from Inforonics.)

For a description of keyboarding specifications to create word processing data for transmission to Inforonics, send for our Technical Memorandum, "A New Way to Create Cataloging Data from Word Processors".

4. Mini-Computer Data Base Input

A better alternative to a word processor, though somewhat more expensive, is a programmed mini-computer. Rather than adding additional software capability to a word processor to make it more versatile, a small, general purpose mini-computer is programmed to handle the data entry task. Inforonics has developed a system which we call a Data Base Input Terminal. This equipment and its associated software extend the data entry concept to achieve field prompting and data validation. Field prompting is the capability of the system to ask the typist for manuscript items by automatically supplying the field tags assigned to each item. The terminal accomplishes this by displaying a "work sheet" tailored to the individual manuscript. This feature allows input from raw source data rather than editing or transcribing data onto clean manuscript or data creation sheets. For monograph cataloging this work sheet will look similar to an OCLC or RLIN screen format. The second function, data validation, is the capability of the system to inform the typist by an audible sound that an error has just been made in the input. The system validates MARC II records catching a majority of the cataloging, typing or editing errors before the document being cataloged leaves the cataloger's hands.

The data base input system also has a printing capability to produce copy for proofreading, an editing capability to correct or change copy, and a transmission capability with error detecting and correction functions. The edit function has the full data validation capability of the keyboard input mode so that errors are also detected while correcting copy.

Including software, the cost of this system ranges from $15,000.00 for a single terminal system to $40,000.00 for a multi-terminal system. It should be remembered that the mini-computer may also perform other data processing functions such as mailing list maintenance and accounting. The cost for data transmitted to Inforonics from this terminal is approximately $0.15 per 1,000 characters (or $0.08 per catalog record). (A complete description of the Inforonics' Data Base Input Terminal can be obtained by writing Inforonics.)

5. Data Entry Using In-House Computer Centers

Customers can use in-house data processing centers to prepare text data using them and deliver data to Inforonics in magnetic tape form. Data is entered into these systems using punch cards, key to diskette, key to disk, key to tape, or an on-line data entry and edit system. Once entered into the computer, data can be printed out, proofread, and edited. Complete and correct files can be written out on magnetic tape and delivered to Inforonics. Two types of formats can be accepted by the Inforonics' system. The preferable format is industry standard

ASCII 1600 BPI magnetic tape recorded in a variable length format with fields tagged, according to the data structure of the manuscript, one physical record to one logical record. There are some conventions for representing character sets extending beyond standard ASCII which need to be followed if the manuscript contains special symbols. A complete description of the ASCII magnetic tape format can be obtained by writing Inforonics, Inc.

The second tape format accepted by Inforonics is industry compatible EBCDIC in which each item in the manuscript is assigned a fixed field position in the record. A description of the EBCDIC format can be obtained by writing Inforonics, Inc.

Magnetic tape from a customer's computer center is a very reliable, low cost method for provision of information to Inforonics as long as the conventions are followed. The cost for this data entry is $0.06 per 1,000 characters (or approximately $0.04 per catalog record).

6. On-Line Data Entry

The last method of data input for Librarians used by Inforonics is text entry on-line to Inforonics' computer. The method allows rapid data entry, immediate printout at the user's terminal, and on-line editing to remove typographic errors or to revise cataloging. Although the system is flexible, it is costly because of the telecommunication and computer resources used. The charges are made on a connect hour basis at approximately $20.00 per hour. This method on the average costs $2.00/1,000 characters.

This on-line input method is rarely used for encoding cataloging data files except for small last minute additions or corrections. In the latter application, the costs are low because a small amount of data is entered. The costs are easily justified because the rapid communication capability of the method greatly speeds up the production process in the final stages of data entry.

SPECIAL INPUT PROBLEMS - CONVERTING BACK FILES

Users have a special problem when converting manual card files to data base supported catalogs. Although a great number of machine form catalog records exist from the Library of Congress and network data bases which can be located and copied to convert a back file, our experience has shown that there may still be a great many unavailable records which need to be keyboarded. The problem of conversion of the unavailable records to machine form has several facets. First, it requires a large amount of keyboard labor compared to that required for the on-going cataloging. Second, the data is rarely in the form needed for input. For example, field tags may not be present, or fields added in the data base design may be missing or appear in a separate section of the catalog such as a shelflist.

Inforonics offers several solutions to this problem: (1) keyboarding may be handled by Inforonics; (2) keyboarding may be performed over-

seas; or (3) a more powerful OCR scanner may be used to read previously typed or typeset cataloging. In keyboarding, the field tagging is done while typing. If scanning is used, field tagging is done by a complex recognition program which processes the scanned output.

1. Keyboarding by Inforonics

For large back files of catalog cards, Inforonics will keyboard the data. Our typists will interpret the data, determine the field tag, and insert it while typing. The data will be printed out, proofread and corrected. For normal card copy, the cost for this input ranges from $3.15 per 1,000 characters to $6.00 per 1,000 characters depending on the complexity of the copy and on whether Inforonics or the customer does the proofreading. For a catalog record this rate results in a cost per entry of $1.80 to $3.00 per record.

2. Overseas Keyboarding

An equivalent service to the above can be provided using overseas keyboarding labor from India, Korea or the West Indies. The costs of this service, including Inforonics supervision costs, are $1.50 per 1,000 characters including proofreading and correction.

3. OCR Scanning

There are several service bureaus which employ Kurzweil scanners (Kurzweil is a subsidiary of Xerox Corporation). This scanner can read typeset or typewritten cataloging with considerable accuracy. The resulting file is computer processed to edit it to a field tagged form. This system is well-suited to simple cataloging only. Complex text requires a human edit pass to detect errors and supply field tags for ambiguous typographic forms read by the scanner. Inforonics will subcontract this service and charge the customer between $0.80 and $1.50 per 1,000 characters depending on the complexity and size of the conversion tasks.

SUMMARY

Beginning with newspaper publishers, more and more publishers are having authors and editors prepare their manuscripts in machine form so that they can be converted directly to type. The number of methods for creating this machine form manuscript are numerous and each has its capabilities and advantages.

The OCR method, while its keyboard equipment is cheapest to acquire, is too error prone to be useful in every-day production. Where keyboards are remote from the scanners, the OCR image impression quality is difficult to control.

Word processors, if already available to an author or editor, are most cost effective for low-volume use. No additional equipment cost is required and they offer reliable, easy input and editing capability.

Pre-programmed mini-computers sold by Inforonics and other systems vendors are best because they can be tailored to make a publisher's particular input keyboarding easy and fast, justifying their cost for high-volume use. Clever users or users with access to programmers can save equipment costs by programming their own micro- or home computers to accomplish a data input task. Finally, if large amounts of backlog data are to be converted, overseas keyboard services offer fast, low-cost capability.